DATA CLUSTERING USING MAXIMUM DEPENDENCY OF ATTRIBUTES AND ITS APPLICATION TO CLUSTER AGRICULTURAL PRODUCTS

HAFIZ BIN KAMAL LEANG

A thesis submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Computer Science (Software Engineering)

FACULTY OF COMPUTER SYSTEM AND SOFTWARE ENGINEERING

UNIVERSITI MALAYSIA PAHANG

MAY, 2012

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

ABSTRACT

This project is about understanding the method of Clustering Data using Rough set Theory. The technique used is Maximum Dependency of attributes. The way this technique work is by calculating the degree of each attribute and selecting the highest dependency based on the degree. The highest degree of attribute will be chosen as the best attribute to be used to cluster the data. A system will be built by using Visual Basic (VB) that will implement this technique to cluster large data faster and easier.

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

ABSTRAK

Projek ini adalah mengenai kajian untuk memahami teknik untuk mengklasifikasikan data menggunakan teori set kasar. Teknik yang digunakan adalah teknik pergantungan maksimum sifat-sifat. Teknik ini digunakan dengan mengira darjah setiap sifat dan seterusnya memilih pergantungan yang paling tinggi berdasarkan darjah yang dikira. Darjah sifat yang paling tinggi akan dipilih sebagai sifat yang paling bagus untuk mengklasifikasikan data. Sebuah sistem akan dibina menggunakan perisian komputer Visual Basic (VB) yang fungsinya untuk melaksanakan teknik ini dalam mengklasifikasikan data yang besar dengan cepat dan senang.

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

TABLE OF CONTENTS

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

LIST OF TABLE

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

LIST OF FIGURES

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

CHAPTER 1

INTRODUCTION

This chapter briefly discuss on the overview of this research. It contains five parts. The first part is background of the research, followed by the problem statement. Next are the objectives where the project goals are determined. After that the scopes of the system and lastly is the thesis organization which briefly describes the structure of this thesis.

1.1 Background

Knowledge discovery is a concept that describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. Also known as deriving knowledge from the input data. Knowledge discovery can be divided into categories based on what kind of data is searched and in what form is the result of the search represented. It is also developed out of the data mining domain, and is closely related to it in terms of methodology and terminology. Knowledge discovery is the most well-known branch of data mining and also known as Knowledge Discovery in Database (KDD). The way it works is, it creates abstractions of the input data. Gained through the process is the knowledge that may become additional data that can be used for further usage and discovery.

Data mining is one of the step in KDD process where data analysis is applied and discovery algorithms that, under certain conditions, produce a particular enumeration of patterns over the data. The data mining component of the KDD process usually involves repeated iterative application of particular data mining

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

method. The methods are classifications, regressions, summarization, dependency modeling, and change and deviation detection. After the general methods of data mining have been outlined, it will then construct specific algorithms to implement these methods. The three primary components that can be identified in any data mining algorithm are model representation, model evaluation, and search.

Clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Classification is a data mining technique used to predict group membership for data instances. For example, classification can be used to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”.

In real life, there are many of type of data that can be collect to be analyzed. When analyzing the data, there are often problems when we want to group the data according to their uniqueness. This often because there is no unique attributes in the data.

There are many types of fruits that can be found in Malaysia. There are so many types of fruits that sometimes not all of them have been seen or ate by a person. Because of this, fruits also have become one of the main sources of income for people living in Malaysia. The reason for why fruits need to be classified is that, when selling fruits, they need to know what attribute that the fruits have and after that separate it into several groups of fruits. This is so that the fruits can graded and sell with a different price.

In this research, the data that have been used are fruits data. The problem from using this data is, it is hard to group the fruits because of no uniqueness in the

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

attributes. To solve this problem, this research will use the maximum dependency of attributes technique to group the fruits data.

1.2 Problem Statement

Having no unique attributes makes it hard to group the data. Thus, another technique is used to cluster the agricultural data.

Rough set is used because this technique able to handle with this kind of data compared to other techniques. Most of the other kind of techniques only can handle numerical data type which is not the kind of data used in this research. Rough set techniques can handle multi-valued data in this research.

1.3 Objectives

The following shows the objectives of the research:

i. To group the mushrooms data according to their dependencies of their attributes
ii. To apply the rough set technique into real life case.

1.4 Scopes

The scopes of this research are shown below:

i. The clustering used maximum dependency of attributes technique.
ii. The used of agricultural data consists of mushrooms.

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

1.5 Thesis Organization

This thesis is organized as follows. Chapter 1 will contain the introduction of this research. Chapter 2 will contain all the literature review that are found for the purpose of doing this research. Chapter 3 consists of the methodology of this research that includes all the technique, algorithm and all the method that are needed to obtain the objectives of this research. Chapter 4 contain the information of the implementation of the application developed based on this research. Chapter 5 will have the conclusions for this research.

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

CHAPTER II

LITERATURE REVIEW

This chapter briefly discusses about the literature review of Agriculture, Knowledge Discovery in Database (KDD), Data Mining, Data Clustering, and Rough Set Theory (RST). The first section is about Agriculture, followed by KDD. After that data mining and data clustering, and lastly Rough Set Theory.

2.1 Agriculture

Agriculture is basically referred to as the cultivation of animals, plants, fungi and other life forms for food, fiber, and other products that are used supply human daily life. Agriculture was the main method in rise of sedentary human civilization, whereby farming of domesticated species created food surpluses that nurtured the development of civilization. Agricultural science is the study of agriculture. Agriculture also includes the observation of certain species of ant and termite, but generally speaking refers to human activities.
(http://en.wikipedia.org/wiki/Agriculture)

Now days, agriculture products were sold using a knowledge-based intelligent e-commerce system. This system will provides products sales, financial analysis and sales forecasting, and not only that it also provides feasible solutions or actions based on the results of rule-based reasoning. This intelligent system will integrates a database, a rule base and a model base to create a tool of which managers can use to deal with decision-making problems using the internet.(U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, 1996).

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

In agricultural production, several types of different methodologies and processes that require a rather high energy input. At the same time, the markets require output products of high quality. The activities can be classified based on the applied methodology, technology and application fields (Hashimoto et al., 2002). These issues appear among the scientific topics of the workshops and conferences organized by the two technical committees, which are the Modeling and Control in Agricultural Processes and Intelligent Control in Agricultural Automation, within the International Federation of Automatic Control (IFAC).(I. Farkas, 2003)

2.1.1 Agriculture in Malaysia

Agriculture in Malaysia contributes up to twelve percent of the nation's Gross Domestic Product (GDP). From the population of Malaysia, sixteen percent are employed through some sort of agriculture. British have established the large-scale plantations. Opportunity has been opened by these plantations, for new crops such as rubber (1876), palm oil (1917), and cocoa (1950). A number of crops are grown for domestic purpose such as bananas, coconuts, durian, pineapple, rice, and rambutan. Productions of exotic produce in Malaysia are mainly because of the proper climate in Malaysia. It is located on a peninsula in Southeast Asia. This area is very rarely affected by hurricanes or drought. Humidity level maintains around ninety percent in Malaysia, it is because of its location close to the equator. The weather stays hot and humid all year round. Malaysia is very populated with hills and large scale agriculture requires a huge amount of flat land. Malaysia does not have a strong temperature climate, because of these disadvantages, Malaysia cannot produce enough rice and other food products to supply the country and forces it to import.
(http://en.wikipedia.org/wiki/Agriculture_in_Malaysia)

In 1999, Malaysia produced 10.55 million metric tons of palm oil, thus making it one of the world's largest producers till today. Almost 85 percent or 8.8 million metric tons of this was exported to international market. Malaysia is one of the world's leading suppliers of rubber, producing 767,000 metric tons of rubber in

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

1999. However, in 1990s, palm oil production is focused more by large plantations companies as it is more profitable. In producing of cocoa, Malaysia has claimed world's fourth-largest with 84,000 metric tons in 1999.

Logging in the tropical rainforest is an important export revenue earner in East Malaysia and in the northern states of Peninsular Malaysia. In 2000, Malaysia produced 21.94 million cubic meters of sawed logs, earning RM1.7 billion from exports. Tropical logs and sawed tropical timber is sold more by Malaysia abroad than any other country, and is one of the biggest exporters of hardwood. Despite attempts at administrative control and strict requirements regarding reforestation in the early 1990s, logging companies often damage the fragile tropical environment. Sharp criticism from local and international environmentalist groups gradually led to bans on the direct export of timber from almost all states, except Sarawak and Sabah. In December 2000, the government and representatives of indigenous and environmentalist groups agreed that there is a need to adopt standards set by the international Forest Stewardship Council (FSC), which certifies that timber comes from well-managed forests and logging companies have to be responsible for reforestation.(http://www.nationsencyclopedia.com/economies/Asia-and-the-Pacific/Malaysia-AGRICULTURE.html)

Malaysia has relied heavily on conventional methods to produce, increase and sustain food productions in the early years of developing the agricultural sector. This is because, a large amount of chemicals fertilizers are needed to supply plant nutrients and chemicals to get rid of pest and diseases. However, in recent years, as a result of increasing awareness on health and environment issues, systematic programs have been introduced to optimize the use of resources on a sustainable basis including the recycling of waste products for food production and environment protection. The successful use of agriculture wastes such as rice, straws and husks, empty oil palm fruit bunches, saw dust, animal droppings, POME etc. and the implementation of good agricultural practices including biological control methods such as IPM are positive steps undertaken to reduce the dependence on chemicals,

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

and to move towards more natural and healthier methods of food production. Integrated and mixed farming is one successful way of optimizing the use of resources for maximizing income.(F. Ahmad, 2001)

2.2 Knowledge Discovery in Databases

Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition/AI, visualization, and high-performance and parallel computing.(U. Fayyad, 1997).

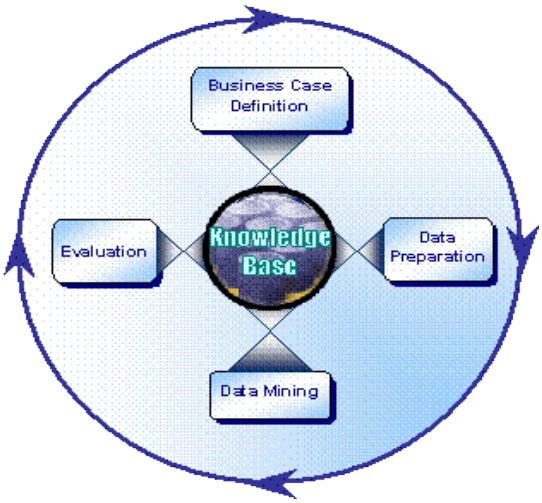
At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data, which are typically too voluminous to understand and digest easily into other forms that might be more compact, for example, a short report, more abstract, for example, a descriptive approximation or model of the process that generated the data, or more useful, for example, a predictive model for estimating the value of future cases. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction. (W. Wen, 2007)

Knowledge Discovery in Databases (KDD) is the automated discovery of patterns and relationships in large databases. Large databases are not uncommon. Cheaper and larger computer storage capabilities have contributed to the proliferation of such databases in a wide range of fields. Scientific instruments can produce terabytes and petabytes of data at rates reaching gigabytes per hour. Point of sale information, government records, medical records and credit card data, are just a few other sources for this information explosion. Not only are there more large databases, but the databases themselves are getting larger. The number of fields in large databases can approach magnitudes of 10^2 to 10^3. Record numbers in these databases approach

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

magnitudes of 10^9. KDD employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization. It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation. KDD has been more formally defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.(Susan P. Imberman, 2001)

2.2.1 KDD Process



KDD begins with the Business Case Definition and proceeds with Data Preparation, Data Mining and Evaluation processes in cyclic order. But the processes are very iterative in nature. Any issues or configuration settings in Data Preparation may result into revisiting and fine-tuning the Business Case Definition. Findings or non-interpretable results from Data Mining process may fall back on the Data Preparation or back to Business Case Definition. Same is the case with Evaluation process.

The Knowledge Base is merely a representation of the database where the business case model, data, metadata, data preparation rules, data mining algorithms, results

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

and evaluation information is kept. It acts as a common pool of information / knowledge, which facilitates the iterations and improves the quality of the model for better results. (Graham J. Williams, Z. Huang, 1996)

The following shows the activities involved in each of the KDD processes.(http://www.executionmih.com/data-mining/kdd-process-preparation-evaluation.php)

Business Case Definition

- Business Goals, Objectives, Critical Success Factors
- High level business cases / issues
- Gap analysis with respect to the current business processes and IT systems
- Framework for the complete Data Mining process

Data Preparation

- Data (as well as Metadata) Quality Analysis
- Data Mining input parameter specification
- Data selection and preparation

Data Mining

- Data Management
- Data Mining Model Build
- Output construction in form of Visualization and Interfaces

Evaluation

- Utilization of data mining output in business processes
- Collection of data from the business processes after data mining
- Assessment / interpretation of Data Mining output

2.2.2 Example of KDD Processes

A model of KDD process from an insurance domain has been taken to make a useful example of KDD process. The insurance company maintains large databases with

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

millions of records to be maintain. Accessible to the customer are the data of transaction oriented, that contain the transactions performed on individual policies. The transaction might be new business, a renewal, a cancellation of a policy, a change to some details of the policy, a claim on a policy, etc. The Source Data is made by the data owners by constructing a relational table and various relational operations from the original data.

The Source Data then transformed into a Working Data by applying a suite of operations on it. The major transformation was from a transaction oriented view of the data to a policy oriented view. Involved in the transformation is an elaborate analysis of the data, which leads to the implementation of a collection of automated operations to perform the task. The important of the automation is so that different transformations could easily be performed on the data as the data became better understood. Figure 1 will describe the actual process with indications of size for a small trial database.

The primary task is encapsulated in the Preprocessing stage. This commenced with the cleansing of the Source Data:

a. Records with missing (critical) values were removed
b. Certain field values were transformed to forms more appropriate for analysis.

The transformations ranged from simple calculations, such as the determination of an age rather than a birth date, to mappings of large range categorical values to a smaller set of categorical values (required in the context of particular data mining tools).

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

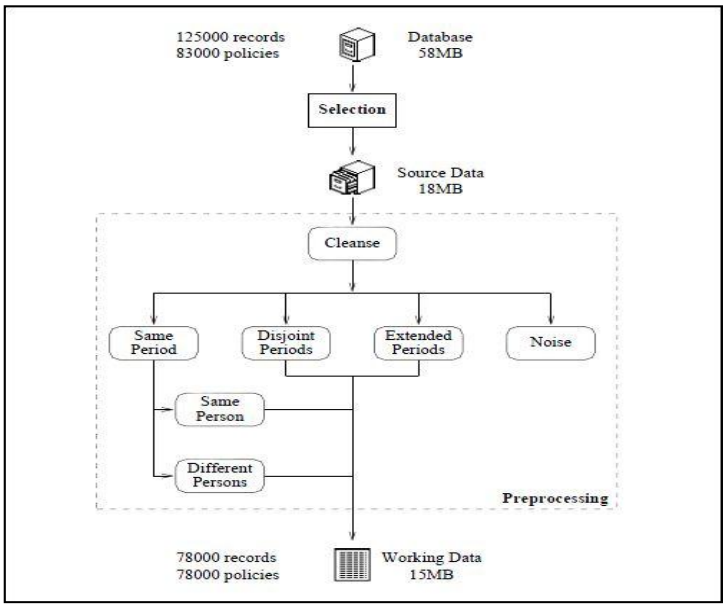


Figure 1: Preprocessing

The major effort expended in the Preprocessing stage was the merging of multiple transactions into single policies. This task required:

- The identification of individual policies
- The merging of multiple transaction policies into single records
- The creation of new fields to record aggregate information

To automate the process of merging, a collection of rules was developed. This facilitated the critical process of revising earlier stage as we iterated through the KDD process. The process was implemented as a collection of filters that could be linked together manually or via a user interface.

Having produced a clean Working Dataset, a task that required considerable effort, the task of most interest to the customer which is exploring the data with a variety of tools, could be addressed. This exploration required revisions to be made to earlier decisions, including the extraction of further attributes from the database and various tuning of the cleansing and merging tasks.

A variety of Data Mining explorations of the data were performed, although many not proving to be particularly insightful, but some leading to interesting snippets of knowledge. StarTree from the Darwin suite of Data Mining tools (Thinking

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

Machines Corporation 1995), for example, was used to build a decision tree to predict if a claim might be made of policy. This analysis identified a number of hot spots in the data which, when combined with further information derived from the data (relating to periods of exposure and size of claim costs, could be used to pinpoint previously unrecognized high risk areas.

The most important element in this KDD exercise was the determination of whether the discovered patters were useful. Subjective opinion lead to the development of some objective criteria for the evaluation of the patters discovered. For example, a discovered rule was deemed to be interesting if it was derived from multiple policies where the total claim cost was significant (above a certain threshold). In determining the worthiness of a rule, extra data not used in the actual Data Mining stage was used, again sometimes requiring modifications to be made to earlier stages of the KDD process.

2.2.3 Application of KDD in computer science fields

Knowledge discovery in database (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. In computer science, KDD is mostly used to manage a large amount of data using data mining.

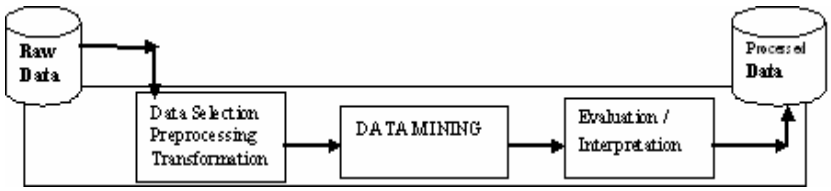


Figure 2: KDD process

Figure 2 shows the process of KDD. It is interactive and iterative involving, more or less, the following steps (R. Kalavathy,_ R.M. Suresh, R. Akhila, 2007):

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

a. Understanding the application domain includes relevant prior knowledge and goals of the application.
b. Extracting the target data set includes selecting a data set or focusing on a subset of variables.
c. Data cleaning and preprocessing includes basic operations, such as noise removal and handling of missing data. Data from real-world sources are often erroneous, incomplete, and inconsistent, perhaps due to operation error or system implementation flaws. Such low quality data needs to be cleaned prior to data mining.
d. Data integration includes integrating multiple, heterogeneous data sources.
e. Data reduction and projection includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods.
f. Choosing the function of data mining includes deciding the purpose of the model derived by the data mining algorithm.
g. Choosing the data mining algorithm(s) includes selecting method(s) to be used for searching patters in data, such as deciding on which model and parameters may be appropriate.
h. Data mining includes searching for patterns of interest in a particular representational form or a set of such representations.
i. Interpretation includes interpreting the discovered patters, as well as the possible visualization of the extracted patterns. One can analyze the patters automatically or semi-automatically to identify the truly interesting or useful patterns for the user.
j. Using discovered knowledge includes incorporating this knowledge into the performance system, taking actions based on knowledge.

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

2.3 Data Mining

Data mining is a knowledge that made offers to new theories, techniques, and tools for processing large volumes of data. Practitioners and researchers have been focusing their attention towards data mining, as evidence by the number of publications, conferences, and application reports. It is also defined as extracting structured information, such as patterns and regularities, from database. Also known as Knowledge discovery in database (KDD), the process is important because it provides means for understanding data, including the generation of predictive rules.

Data mining actual task is the automatic or semi-automatic analysis of large quantities of data in order to extract previously unknown interesting patters such as groups of data records, unusual records, and dependencies. These patterns can then be seen as a kind of summary of the input data, and used in further analysis or for example in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation nor result interpretation and reporting are part of the data mining step, but do belong to the overall data mining process as additional steps.

Data mining is not a brute-force crunching of bulk data, blind application of algorithms, going to find relationships where none exist, presenting data in different ways, a database intensive task, and a difficult to understand technology requiring an advance degree in computer science, but it is a hot buzzword for a class of techniques that find patters in data, a user-centric, interactive process which leverages analysis technologies and computing power, a group of technique that find relationships that have not previously been discovered, not reliant on an existing database, and a relatively easy task that requires knowledge of the business problem or subject matter expertise.

To conduct an effective data mining, the first step to take is to examine what kind of features an applied knowledge discovery system is expected to have and what kind

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

of challenges one may face at the development of data mining techniques. List of them are as follows:

a. Handling of different types of data.
b. Efficiency and scalability of data mining algorithms.
c. Usefulness, certainty and expressiveness of data mining results.
d. Expression of various kinds of data mining results.
e. Interactive mining knowledge at multiple abstraction levels.
f. Mining information from different sources of data.
g. Protection of privacy and data security.

There have been many advances on researches and developments of data mining, and many data mining techniques and systems have recently been developed. Different classification schemes can be used to categorize data mining methods and systems based on the kinds of databases to be studied, the kinds of knowledge to be discovered, and the kinds of techniques to be utilized, as shown below:

a. What type of databases to work on.
b. What type of knowledge to be mined.
c. What type of techniques to be utilized.

Methods for mining different kinds of knowledge, including association rules, characterization, classification, clustering etc. are examined in depth. For mining a particular kind of knowledge, different approaches, such as machine learning approach, statistical approach, and large database-oriented approach, are compared, with an emphasis on the database issues, such as efficiency and scalability.

2.3.1 Example of Data Mining

In KDD process, the model of association rules is important and the most representative association rules algorithm is Apriri algorithm. The objective of association rules mining is to fast discover the interesting association or related

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional

relationship between attributes of the mass data in large-scale database. Since the association rules extraction has something to do with the source database system, then in a sense, the corresponding association rules are not generated by the direct use of DBS but by certain transformation. Therefore, in order to increase the scanning speed of the large-scale DBS and extract the association rules quickly, we must change the quantity related problems into logical related problems.

Table 1 gives a simple database example. The table mainly shows the potential associations among higher education, wages, sex, teacher status and age, and shows the future tendencies which may be lead.

Table 1: A simple example of database

RECID	SEX	AGE	KNOWLEDGE	OCCUPATION	WAGES
100	Male	46	Doctor	Teacher	7500
200	Female	32	Master	Teacher	6500
300	Male	35	Bachelor	Technician	4900
400	Male	40	Master	Teacher	6000
500	Male	37	Doctor	Teacher	7000
600	Male	25	Bachelor	Technician	4000

We do dualization for sex SEX (1:Male, 2:Female); discretization for age AGE (Age ≥ 40,3:middle; Age < 40,4:young); discretization for whether received postgraduate education KNOWLEDGE (master or doctor degree, 5: high; undergraduate diploma or low than undergraduate diploma, 6: low); dualization for occupation (college teachers, 7: teacher; not college teachers, 8:technician); dualization for wages WAGES (average monthly income is higher than 5000, 9: wages> 5000; 10: wages<

Created with nitroPDF professional
download the free trial online at nitropdf.com/professional